DOCUMENT RESUME

ED 390 897 TM 024 179

AUTHOR	DeMauro, Gerald E.; Powers, Donald E.
TITLE	Logical Consistency of the Angoff Method of Standard Setting.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	RR-93-26
PUB DATE	May 93
NOTE	18p.; Version of a paper presented at the Annual Meeting of the National Council on Measurement in Education (Boston, MA, April 17-19, 1990).
PUB TYPE	Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Certification; *Cutting Scores; Judges; *Licensing Examinations (Professions); *Logic; *Pass Fail Grading; *School Psychologists; Standards
IDENTIFIERS	*Angoff Methods; Consistency (Behavior); NTE Specialty Area Tests; *Standard Setting

ABSTRACT

Standard setting on licensure and certification tests is difficult both to execute and to defend. There may, however, be certain minimum standards for standard setting on which most will be able to agree. One such standard is logical consistency. M. T. Kane (1984, 1986) has suggested an approach to evaluating the logical consistency of one widely used method to set passing scores--the Angoff procedure (W. H. Angoff, 1971). This approach is applied here to the standard setting data from a study of the NTE Specialty Area Test for School Psychology. In brief, the Angoff procedure involves obtaining judges estimates of the probability with which minimally competent examinees can be expected to answer correctly each item in a test. For this study, 130 items from the test were used with 19 panelists who were members of the National Association of School Psychologists. The study provided modest evidence that the Angoff procedure does yield results that display a relatively high degree of logical consistency, especially as judged from the mean estimates provided by a panel of judges. (Contains 3 tables and 15 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

RR-93-26

ED 390 897

RESEARCH REPORT

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
H. I. BRAUN
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

LOGICAL CONSISTENCY OF THE ANGOFF METHOD OF STANDARD SETTING

Gerald E. DeMauro
Donald E. Powers

Educational Testing Service
Princeton, New Jersey
May 1993

BEST COPY AVAILABLE

TM024179

Logical Consistency of the Angoff Method of Standard Setting

Gerald E. DeMauro

Donald E. Powers

Educational Testing Service
Princeton, New Jersey 08541

Copyright © 1993. Educational Testing Service. All rights reserved.

Abstract

Standard setting on licensure and certification tests is difficult both to execute and to defend. There may, however, be certain minimum standards for standard setting on which most everyone may be able to agree. One such standard is logical consistency.

Kane (1984, 1986) has suggested an approach to evaluating the logical consistency of one widely used method to set passing scores—the Angoff procedure (Angoff, 1971). This approach is applied here to the standard setting data obtained in a study of the NTE Specialty Area Test for School Psychology.

Logical Consistency of the Angoff Method of Standard Setting[1]

Opinions differ considerably regarding which methods of standard setting yield the most defensible results for educational and psychological tests. As Berk (1986) has noted, the process of setting standards on standardized tests is easily the most complicated technical issue in criterion-referenced measurement. Despite considerable research, standard setting remains "controversial to discuss, difficult to execute, and almost impossible to defend" (p. 565). With respect to licensure tests, some critics have expressed the opinion that the way in which standards have been established is the single most serious weakness of such tests. According to Haertel (1987), for example, currently used procedures seem to require judges to speculate about the test performances of hypothetical, minimally competent persons, and the resulting numbers may therefore lack any significant meaning.

Furthermore, it is well documented (Berk, 1986) and generally acknowledged, e.g., in the Standards for Educational and Psychological Testing, (AERA, APA, NCME, 1985), that large discrepancies can exist among the passing scores produced by different methods. Some researchers have noted, however, that there appears to be an increasing agreement that a procedure discussed by Angoff (1971)..."produces more reasonable standards than do its competitors" (Jaeger, 1988, p. 17).[2]

The Standards are clear on the need for careful study of the reliability and validity of decisions and inferences based on cut scores. However, there appears to be relatively little relevant evidence here. For teacher certification tests in particular, few if any of the many state-conducted validity studies have been designed to assess the accuracy or validity of decisions

---

[1] This paper is based on a presentation given at the 1990 NCME Annual Meeting, April 17-19, in Boston.

[2] Angoff (1971) has frequently credited Dr. Ledyard Tucker with originally recommending the procedure discussed in his 1971 chapter. Therefore, the procedure has appropriately been referred to as the Tucker/Angoff method. Throughout this paper, however, we have used the term "Angoff procedure."

based on cut scores (Haertel, 1987; Madaus, 1987). This is true even though the appropriate way to compute the reliability of a standard setting procedure is readily apparent—at least in theory. The relevant index is the extent to which the method produces consistent classifications of an examinee (as competent or incompetent) when applied to (a) different samples of items, (b) different samples of judges, or (c) different occasions of judgment (Jaeger, 1988).

Logical consistency is an even more fundamental characteristic than classification consistency. That is, at a minimum the results of a particular procedure should be consistent with the logic that underlies it. Plake, Melican, and Mills (1991) have discussed some of the factors that may influence intrajudge consistency. These factors include characteristics of the judges and of the test items. Melican and Thomas (1984) focused on identifying characteristics of test items that may influence the quality of judgments. Bejar (1983) suggested that estimating item difficulty is not an easy task, even for experts.

Van derLinden (1982) proposed using IRT methodology to evaluate the degree to which judgments correspond with an underlying model. Kane (1984, 1986) has suggested another way in which the logical consistency of one widely used standard setting procedure, the Angoff method, could be evaluated. According to the assumptions underlying the Angoff method, examinees who score at or just above the passing score that is set by the method can be regarded as minimally competent. The proportions of these particular test takers who do in fact actually answer each item correctly should correspond to the estimates provided by the standard-setting panel. As Kane has suggested, inconsistency can be defined as the extent to which these two sets of proportions differ.

The major objective of the study reported here was to assess the extent to which the Angoff procedure yields results that are, as defined above, logically consistent. This objective was investigated within the context of the ETS Teacher Programs, which offer a variety of

teacher certification tests. Passing scores on these tests are typically established by state agencies, usually on the basis of the Angoff procedure.

Methods

Briefly, the Angoff procedure involves obtaining judges' estimates of the probability with which minimally competent examinees can be expected to answer correctly each item in a test. (In the standard setting study that generated the data examined here, judges were asked to select one of the following pre-specified percentages--2, 10, 25, 40, 60, 80, or 98.) These estimates are then summed over all items to yield a suggested passing score for the test.

For this study, standard-setting data were obtained from a validity study of the NTE Specialty Area Test for School Psychology, a 135-item test that is used by states to certify school psychologists. Of the 135 items on the test, 130 were scored. Raw score statistics based on 6,373 test takers were as follows:

Mean = 85

SD = 15

Median = 87

Range = 25 to 122

Skewness = -.6

Kurtosis = .3

In this validity study, 19 panelists who represented the membership of the National Association of School Psychologists (NASP) met in November 1988 to help determine a passing score that could be adopted by certifying agencies. Panelists were selected by NASP to represent its membership with regard to gender (10 female, 9 male) and professional responsibilities. Ten panelists identified themselves as college- or university-based trainers of

school psychologists, six as practitioners in elementary or secondary schools, and three as holding some other position.

Test score data from a recent administration of the test were used in order to determine the correspondence between Angoff probabilities derived from the panel judgments and the actual percentages of minimally competent examinees (as determined by the passing score) who answered each item correctly. Specifically, consistency was checked in two ways. First, passing scores were computed for each judge by summing the Angoff probabilities assigned to each test item by the judge. Next, for each judge, examinees were identified who scored within half a standard deviation of the passing score set by the judge. This range was necessary in order to ensure that a sufficient number of examinees were classified as minimally competent. (For the judges in this study the number of examinees in this range varied from 245 to 2,683. For all but two judges the number exceeded 1,000). The percentage of these minimally competent examinees who actually answered each item correctly was then determined. These actual percentages, as well as the Angoff probabilities, were converted to delta values, a normalized transformation of the percentage of examinees answering each item correctly (delta = 13 - 4z, where z is the value of the normal curve corresponding to percentage correct). A correlation was then computed between judges' Angoff probabilities and the actual difficulty of an item for minimally competent examinees, i.e., those who fell within half a standard deviation of each judge's passing score. This procedure was repeated for each of the 130 items that was scored for the test.

Consistency of judgments was checked within items. Correlations were calculated over items between the mean (over all judges) of an item's Angoff probability and the mean (over all judges) of minimally competent examinees' actual performance on the item. Because, in practice, passing scores are usually determined from the mean estimates of a panel of judges,

this index may be the most appropriate one (of those considered here) for evaluating logical consistency.

Results

Correlations Between Estimated and Actual Probabilities

Table 1 shows a substantial variation in correlations (over judges) between Angoff probabilities and actual item difficulties for the 130 items. The median correlation was .44. Correlations (over items) between estimated and actual values for 19 judges ranged from .25 to .56. The median was .42.

Mean Differences Between Actual and Judged Difficulty (over all items) for Judges

Table 2 reveals that judges varied considerably in their estimates of average item difficulty for minimally competent examinees (from a mean delta of 9.65 to 14.44). Relatively little information was available, however, about possibly relevant characteristics of the judges to help explain the differences between judges with respect to the level of their estimates.

The directions of the average discrepancies between actual item difficulties and judged item difficulties are worth noting also. Most (15) of the 19 judges' estimates of the difficulty of items for minimally competent examinees were on average higher than the actual mean difficulty. That is, examinees who were classified as minimally competent more often found the items somewhat easier than they were judged to be. There was no apparent relationship between the degree of under-(over-) estimation of item difficulty and item content (as indexed by sic content categories).

Two estimates of judge consistency are given in Table 3. The first is the sum of absolute differences (in delta units) between estimated and observed item difficulties for each judge over the 130 items. Table 3 also shows the correlation over items between estimated and observed item difficulties for each judge. A large absolute difference between estimated and observed

item difficulties and a high correlation between these two measures is not likely, because minimally competent examinee groups were sampled on the basis of summed item difficulty estimates of the judges. Rather, a large absolute difference indicates that a judge was inconsistent, underestimating the difficulty on some items and overestimating the difficulty on others. Such inconsistency would also be reflected in lower correlations between observed and estimated difficulties. Clearly, some judges were more consistent than others.

The correlation over items between the mean (averaged over judges) of judges' Angoff probabilities and the mean (averaged over judges) of minimally competent examinees' actual performance was .71 ($p < .001$). Thus, as a panel, judges were reasonably consistent.

Discussion

The study provides modest evidence to suggest that the Angoff procedure, one of the most widely used methods of setting standards on certification and licensure tests, including those used to certify teachers, does yield results that display a relatively high degree of logical consistency, especially as judged from the mean estimates provided by a panel of judges. As expected, perfect consistency was not found, and experience with the kind of evidence generated here will be needed in order to establish what constitutes adequate agreement (Kane, 1984). Data of this type do, as Kane (1986, 1987) has contended, appear to have potential for providing at least a partial basis for choosing among standard setting methods. For the data evaluated here, the Angoff procedure seems to meet the criterion of logical consistency—a (perhaps) necessary condition for the validity of the method—reasonably well. This result should be viewed in light of previous findings that, typically, judges do not agree especially well with respect to their ratings of test items' absolute difficulties (e.g., Brennan & Lockwood, 1980; Skakun & Kling, 1980), even though they can agree about the relative difficulty of items (Thorndike, 1982). Moreover, this finding is consistent with the limited empirical evidence on the comparative

reliability of standard setting procedures, which suggests that the Angoff procedure frequently yields cut scores that are more reliable than those produced by other methods (Jaeger, 1988).

Another practical implication of the results is that a suggestion by Kane (1986)—for maximizing agreement between Angoff estimates and actual item probabilities at the passing score—appears to be feasible. The suggested procedure entails setting the score such that the sum of absolute discrepancies between actual probabilities and estimated probabilities for all items is minimized. All in all, the findings provide modest support for the continued use of the Angoff procedure to establish standards on certification and licensure tests. The methods used here do, however, need to be applied to other tests in order to establish the generalizability of the results, and of course standard setting procedures should also be subjected to other more stringent criteria.

References

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational Measurement (pp. 508-600). Washington, DC: American Council on Education.

American Education Research Association/American Psychological Association/National Council on Measurement in Education (1985). Standards for educational and psychological testing. Washington, DC: American Psychological Association.

Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. Review of Educational Research, 56, 137-172.

Brennan, R. L., & Lockwood, R. E. (1980). A comparison of the Nedelsky and Angoff cutting score procedures using generalizability theory. Applied Psychological Measurement, 4, 219-240.

Haertel, E. (1987, June). Validity of teacher licensure and teacher education admissions tests. Paper prepared for the National Education Association and Chief State Officers.

Jaeger, R. M. (1988, August). Establishing standards on tests used for certification of education personnel: Validity issues. Paper presented at the Annual Convention of the American Psychological Association, Atlanta, GA.

Kane, M.T. (1984, April). Strategies in validating licensure examinations. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Kane, M. T. (1986). The interpretability of passing scores (ACT Technical Bulletin Number 52). Iowa City, Iowa: The American College Testing Program.

Kane, M. T. (1987). On the use of IRT models with judgmental standards setting procedures. Journal of Educational Measurement, 24, 333-345.

Madaus, G. (1987, October). Legal and professional issues in teacher certification testing: A psychometric snark hunt. Paper presented at the Fifth Annual Buros-Nebraska Symposium on Measurement and Testing, Lincoln, Nebraska.

Melican, G. J. & Thomas, N. (1984, April). Identification of items that are hard to rate using Angoff's standard setting method. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Plake, B. S., Melican, G. J., & Mills, C. N. (1991). Factors influencing intrajudge consistency during standard setting. Educational Measurement: Issues and Practice, 10, 15-22, 22, 25-26.

Skakun, E. N., & Kling, S. (1980). Comparability of methods for setting standards. Journal of Educational Measurement, 17, 229-235.

Thorndike, R. L. (1982). Item and score conversion by pooled judgment. In P. W. Holland and D. B. Rubin (Eds.), Test Equating (pp. 309-318). New York: Academic Press.

Van derLinden, W. J. (1982). A latent trait method for determining intrajudge inconsistency in the Angoff and Nedelsky techniques of standard setting. Journal of Educational Measurement, 19, 295-308.

Acknowledgments

We thank Gerald Melican and Michael Zieky for helpful reviews of an earlier draft.

Table 1

Distribution of Product Moment Correlations between Estimated and Observed Item Difficulties, for Borderline Examinees (N = 130 items)

Correlation Range	Frequency	Percent	Cum. Percent
-.19 to -.15	2	2	2
-.14 to -.10	3	2	4
-.09 to -.05	2	2	5
-.04 to -.01	1	1	6
.00 to .04	2	2	8
.05 to .09	2	2	9
.10 to .14	6	5	14
.15 to .19	8	6	20
.20 to .24	5	4	24
.25 to .29	7	5	29
.30 to .34	7	5	35
.35 to .39	13	10	45
.40 to .44	9	7	52
.45 to .49	16	12	65
.50 to .54	12	9	74
.55 to .59	9	7	81
.60 to .64	11	8	89
.65 to .69	4	3	92
.70 to .74	4	3	95
.75 to .79	7	5	100

Table 2

Mean Estimated and Observed Item Difficulties, in Deltas,

for Borderline Examinees, by Judge

Judge	Estimated Difficulty	Observed Difficulty	Difference
1	11.15	10.95	0.21
2	12.58	12.17	0.41
3	9.90	9.93	-0.03
4	12.02	11.53	0.49
5	10.93	10.85	0.08
6	12.58	12.32	0.26
7	12.31	11.95	0.36
8	11.96	11.61	0.35
9	14.44	13.63	0.81
10	10.16	10.42	-0.26
11	10.96	10.85	0.11
12	9.65	9.93	-0.28
13	12.88	12.23	0.65
14	12.25	11.95	0.30
15	10.14	10.42	-0.28
16	11.45	11.28	0.17
17	10.45	10.33	0.12
18	13.19	12.80	0.39
19	12.80	12.23	0.57
Mean	**11.67**	**11.44**	**0.23**

Table 3

Correlations and Absolute Differences between Estimated and Observed Item Difficulties, in Deltas, over Items, by Judge

Judge	Absolute Difference in Observed and Estimated Difficulties	Correlation between Observed and Estimated Difficulties
1	2.06	0.54
2	2.21	0.45
3	2.65	0.36
4	1.99	0.49
5	2.75	0.40
6	3.66	0.46
7	2.44	0.38
8	2.20	0.52
9	2.26	0.33
10	3.54	0.25
11	2.32	0.52
12	2.77	0.42
13	2.63	0.50
14	2.96	0.35
15	3.16	0.37
16	2.75	0.40
17	2.10	0.56
18	3.36	0.44
19	2.12	0.35
Mean	**2.55**	**0.42**